

ЛЬВІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ІВАНА ФРАНКА

На правах рукопису

Волощук Орест Романович

УДК 004.853+004.855.5

МЕТОДИ КЛАСТЕРИЗАЦІЇ НА ВЕЛИКИХ МАСИВАХ ДАНИХ

01.05.03 — математичне та програмне забезпечення обчислювальних
машин і систем

Магістерська робота

Науковий керівник

Годич Олесь Васильович,

кандидат фізико-математичних наук, доцент

Львів — 2011

ЗМІСТ

Вступ	3
Розділ 1. Огляд стану проблеми та основні поняття	4

ВСТУП

Актуальність теми. Сьогодні все частіше виникають задачі, так чи інакше пов'язані із розбиттям масиву об'єктів на групи за певними критеріями. Із розвитком обчислювальної техніки збільшуються також об'єми баз даних, що можуть піддаватись такому аналізу. Виникає необхідність створення оптимальних алгоритмів обробки великих масивів даних.

Мета і завдання дослідження. Метою дослідження є розвиток методики кластеризації великих масивів даних. Для досягнення цієї мети були сформульовані та вирішені такі основні завдання:

- провести детальний аналіз та дослідити ефективність різноманітних алгоритмів кластеризації;
- здійснити контроль якості результатів роботи алгоритмів;
- розробити методи та виявити можливі оптимізації, що дозволять прискорити виконання задачі кластеризації

Об'єкт дослідження. Об'єктом дослідження є методи ієрархічні та плоскі алгоритми кластеризації даних.

РОЗДІЛ 1

ОГЛЯД СТАНУ ПРОБЛЕМИ ТА ОСНОВНІ ПОНЯТТЯ

Кластеризацією називається процес розбиття певної заданої множини об'єктів на підмножини, що називаються *кластерами*, так, щоб в одній підмножині (кластері) опинились об'єкти, якимось чином схожі між собою. Ці об'єкти також називають *спостереженнями*. Кожне спостереження характеризується певним скінченним набором атрибутів. Якщо у кожного об'єкта є точно n атрибутів, і якщо всі ці атрибути у якийсь спосіб приведено до числового вигляду, то можна розцінювати такий масив даних як набір точок у n -вимірному просторі.

1.0.1. Задачі. В наш час з'явилась низка задач, розв'язання яких на певному етапі вимагає кластеризації вхідних даних. Ці задачі походять із різноманітних сфер людської діяльності – соціології, медицини, математики, і навіть із комерційної діяльності. Кластеризація зазвичай застосовується при розпізнаванні образів, виявленні прихованих закономірностей даних, та ін.

1.0.2. Історія. Техніки кластеризації були практично недоступні до середини ХХ століття. До появи достатньо потужних ЕОМ лише людина могла теоретично здійснити кластеризацію певного масиву даних. На таку ручну кластеризацію накладались би очевидні обмеження – зокрема, неможливість працювати із просторами розмірністю більшою ніж 2.